

Degrees of Freedom of the Reduced Rank Regression

Ashin Mukherjee, Naisyin Wang and Ji Zhu

Department of Statistics, University of Michigan

October 10, 2012

Abstract

In this paper we study the effective degrees of freedom of the reduced rank regression estimator in the framework of Stein's unbiased risk estimation (SURE). We derive a finite-sample exact unbiased estimator of the degrees of freedom for the reduced rank regression. We show that it is significantly different from the number of free parameters in the model, which is often taken as a heuristic estimate of the degrees of freedom for the reduced rank regression. Using the exact unbiased estimator of the degrees of freedom, one can easily employ various model selection criteria such as Mallows' C_p or GCV to efficiently choose an optimal rank for the reduced rank regression problem, which often outperforms its heuristic counterpart in terms of prediction accuracy and successfully avoids computationally expensive data perturbation or bootstrap based methods. We have also extended the proposed approach to other related estimation procedures, including the reduced rank ridge regression and a weighted nuclear norm penalized multivariate regression.

KEY WORDS: Degrees of freedom, Model selection, Multivariate regression, Reduced rank regression.

1 Introduction

Multivariate linear regression is the extension of the classical univariate regression model to the case where we have $q(> 1)$ responses and p predictors. It is commonly used in bioinformatics, chemometrics, econometrics, and other quantitative fields where one is interested in predicting several responses simultaneously.

We can express the multivariate linear regression model in matrix notation as follows. Let \mathbf{X} denote the $n \times p$ predictor or design matrix, with the i -th row $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$. Similarly the $n \times q$ dimensional response matrix is denoted by \mathbf{Y} , where the i -th row is $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iq}) \in \mathbb{R}^q$. The regression parameters are given by the coefficient matrix \mathbf{B} which is of dimension $p \times q$. Note that the k -th column of \mathbf{B} is the regression coefficient vector for regressing the k -th response on the predictors. Let \mathbf{E} denote the $n \times q$ random error matrix with independent entries with mean zero and variance σ^2 . Then the multivariate linear regression model is given by

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}. \quad (1)$$

Note that, this reduces to the classical univariate regression model when $q = 1$. For notational simplicity, we assume that the responses and the predictors are centered, and hence the intercept term can be omitted without any loss of generality. The ordinary least squares approach of estimating \mathbf{B} leads to

$$\hat{\mathbf{B}}_{\text{ols}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

The ordinary least squares estimate amounts to performing q separate univariate regressions and completely ignores the multivariate aspect of the problem, where many of the responses might be highly correlated and hence the effective dimensionality can be much smaller than q . Quite a large number of methods have been proposed in the literature to overcome these drawbacks. Many of them would fall under the general class of *linear factor regression*, where the responses are regressed against a small number of linear combination of predictors commonly known as factors. Examples include principal component regression (Massy, 1965), partial least squares (Wold, 1975), canonical correlation analysis (Hotelling, 1935) and so on. The methods differ in the way they choose the factors. Recently Witten et al. (2009) introduced a penalized canonical correlation analysis using sparse matrix factorization that leads to more interpretable factors and is more suitable for high-dimensional problems. Breiman and Friedman (1997) proposed the curds and whey (C&W) approach which borrows strength by performing a second round of regression of the responses on the ordinary least

squares estimators. The authors also show some close connections of the C&W approach with canonical correlation analysis.

Several penalization methods have been proposed in recent years to address the issues of prediction performance and variable selection in multivariate regression. Turlach et al. (2005) introduced an ℓ_∞ penalty on the rows of \mathbf{B} to encourage simultaneous variable selection. Peng et al. (2009) used a combined penalty function of the form, $\mathcal{J}(\mathbf{B}) = \lambda_1 \sum_{j=1}^p \|\mathbf{B}_j\|_1 + \lambda_2 \sum_{j=1}^p \|\mathbf{B}_j\|_2$ to identify “master predictors” in genomics studies. Note that, \mathbf{B}_j denotes the j -th row of the coefficient matrix. The first part of the penalty imposes sparsity on entries of \mathbf{B} , whereas the second part forces some of the entire rows of $\hat{\mathbf{B}}$ to be zero, encouraging the selection of “master predictors” that influence many response variables. Obozinski et al. (2011) developed asymptotic theory for the ℓ_1/ℓ_2 penalized multivariate regression problem, which can also be thought of as a special case of the joint penalty employed by Peng et al. (2009) with $\lambda_1 = 0$. In particular, they prove that the multivariate group lasso penalty recovers the correct row support with high probability in high-dimensional settings.

The methods discussed above do not account for correlated errors. This motivated Rothman et al. (2009) to propose a joint method for variable selection and covariance estimation. Let $\mathbf{\Omega}$ denote the inverse covariance matrix for the errors. Then the penalty function employed by Rothman et al. (2009) takes the form, $\mathcal{J}(\mathbf{B}, \mathbf{\Omega}) = \lambda_1 \sum_{j \neq j'} |\omega_{jj'}| + \lambda_2 \sum_{j=1}^p \sum_{k=1}^q |b_{jk}|$. This leads to interpretable models that predict better. Lee and Liu (2012) explore this further by introducing two plug-in estimators as well as a joint method that employs weighted ℓ_1 penalty to estimate $(\mathbf{B}, \mathbf{\Omega})$ simultaneously. In addition they prove that their estimation methods are consistent and admit asymptotic distributions.

Yet another line of research focuses on the rank of the regression coefficient matrix. Anderson (1951) proposed a class of regression models that restrict the rank of the true coefficient matrix to be much smaller than the dimensionality of \mathbf{B} , i.e. $\text{rank}(\mathbf{B}) \leq r \leq \min\{p, q\}$. This is a quite reasonable assumption in many multivariate regression problems, which can be reformulated as follows: the q responses are related to the p predictors only through r effective linear factors. It results in the following optimization problem

$$\hat{\mathbf{B}}(r) = \underset{\{\mathbf{B}: \text{rank}(\mathbf{B}) \leq r\}}{\text{argmin}} \quad \|\mathbf{Y} - \mathbf{XB}\|_F^2. \quad (2)$$

Even though the rank penalty makes it a non-convex optimization problem, it admits a closed form solution as we shall see later. Izenman (1975) introduced the term reduced rank

regression for this class of models and derived the asymptotic distributions and confidence intervals for reduced rank regression estimators. A non-exhaustive list of notable work includes Rao (1978), Davies and Tso (1982), Anderson (1999, 2002); see Reinsel and Velu (1998) or Izenman (2008) for a more comprehensive account. Recently, there has been a revival of interest in the reduced rank methods. Instead of restricting the rank, Yuan et al. (2007) proposed to put an ℓ_1 penalty on the singular values of \mathbf{B} also known as the nuclear norm. This can be considered as the convex relaxation to the reduced rank regression criterion, but the resulting optimization problem is quite non-trivial. Bunea et al. (2011) came up with an innovative ℓ_0 -penalized viewpoint of the reduced rank regression. Under that framework they are able to characterize the choice of tuning parameter, which guarantees asymptotic consistency in terms of rank selection. Chen et al. (2012a) adopted sparsity penalties on singular vectors for reduced rank regression problems that lead to more interpretable models. Very recently Chen et al. (2012b) proposed a weighted nuclear norm penalty on the signal matrix \mathbf{XB} ; the resulting optimization problem admits a closed form solution, which enjoys many desirable theoretical properties. We note that most of the theoretical results obtained so far are asymptotic in nature.

In this paper we study the degrees of freedom of the reduced rank regression model. The degrees of freedom is a very familiar and one of the most widely used terms in statistics. We utilize it from ANOVA t-tests to model selection criteria such as AIC and BIC. However, it has been largely overlooked in the reduced rank regression literature except for some heuristic suggestions (Davies and Tso, 1982; Reinsel and Velu, 1998; Chen et al., 2012b). For example, the number of free parameters in a $p \times q$ matrix of rank r , given by $r(p + q - r)$ has been suggested as a naive estimate of the degrees of freedom of the reduced rank regression model restricted to rank $r \leq \min\{p, q\}$. In this paper, we aim to find a finite-sample unbiased estimator of the degrees of freedom for the reduced rank regression model and investigate its properties. The result covers a significant gap in the literature, as the previously suggested naive estimate lacks both statistical motivation and practical performance.

In a nutshell, the degrees of freedom quantifies the complexity of a statistical modeling procedure (Hastie and Tibshirani, 1990). In the case of the univariate linear regression model, it is well-known that the degrees of freedom is the number of estimated parameters, p . However, in general there is no exact correspondence between the degrees of freedom and the number of free parameters in the model (Ye, 1998). For example, in the best subset selection for univariate regression (Hocking and Leslie, 1967) we search for the best model of size $p_0 \in \{1, 2, \dots, p\}$ that minimizes the residual sum of squares. The resulting model has p_0

parameters but intuitively the degrees of freedom would be higher than p_0 since the search for the “optimal” subset of size p_0 increases model complexity (Hastie et al., 2009). In other words, for best subset selection the optimal p_0 -dimensional subspace that minimizes the residual sum of squares clearly depends on \mathbf{y} . Thus the final estimator is highly non-linear in \mathbf{y} , which results in the loss of correspondence between degrees of freedom and the number of parameters in the model.

Similar arguments also apply to the reduced rank regression. Instead of searching for best p_0 -variables as in the case of best subset selection, here we are searching for best q_0 linear combinations of the predictors that minimize the least squares loss, which should intuitively suggest increased model complexity. Since the optimal rank q_0 -subspace depends on the response matrix \mathbf{Y} , the natural correspondence between number of free parameters and degrees of freedom need not hold. This is where reduced rank regression is different from other linear factor regression methods, e.g. principal component regression (Massy, 1965). In principal component regression, the factors are principal components of the design matrix \mathbf{X} , which do not depend on the response \mathbf{Y} , thus the final estimator is still linear in \mathbf{Y} .

The rest of the paper is organized as follows. In section 2, we review the degrees of freedom in the framework of Stein’s unbiased risk estimation (Stein, 1981). Section 3 contains our proposed estimator of the degrees of freedom of the reduced rank regression model. In section 4, we extend the methodology to other related methods, including the reduced rank ridge regression and a reduced rank regression model that uses the weighted nuclear norm penalty. In section 5, we show that the exact unbiased estimator of the degrees of freedom of the reduced rank regression model can be significantly different from the naive estimator through several numerical examples. We also show that when using the degrees of freedom derived from the previous sections, one can gain prediction accuracy over its heuristic counterpart with the reduced rank regression. In section 6, we apply the developed method to a genetic association data example, and we conclude the paper with a discussion in section 7.

2 Degrees of freedom

Stein (1981) in his theory of unbiased risk estimation (SURE) first introduced a rigorous definition of the degrees of freedom of a statistical estimation procedure. Later Efron (2004) showed that Stein’s treatment can be considered as a special case of a more general notion under the assumption of Gaussianity. Assume that we have data of the form $(\mathbf{y}_{n \times 1}, \mathbf{X}_{n \times p})$. Given \mathbf{X} , the response originates from the following model $\mathbf{y} \sim (\boldsymbol{\mu}, \sigma^2 \mathbf{I})$, where $\boldsymbol{\mu}$ is the true

mean which can be a function of \mathbf{X} , and σ^2 is the common variance. Then for any estimation procedure $m(\cdot)$ with fitted values $\hat{\boldsymbol{\mu}} = m(\mathbf{X}, \mathbf{y})$, the degrees of freedom of $m(\cdot)$ is defined as

$$df(m) = \sum_{i=1}^n cov(\hat{\mu}_i, y_i) / \sigma^2. \quad (3)$$

The rationale is that more complex models would try to fit the data better, and hence the covariance between observed and fitted pairs would be higher. This expression is not directly observable except for certain simple cases, for example, when $m(\mathbf{y}) = \mathbf{S}\mathbf{y}$, a linear smoother. In that case, it is not difficult to see that $df(m) = tr(\mathbf{S})$. Stein was able to overcome this hurdle for a special case when $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 I)$. Using a simple equality for the Gaussian distribution, he proved that as long as the partial derivatives $\partial \hat{\mu}_i / \partial y_i$ exists almost everywhere for all $i \in \{1, 2, \dots, n\}$, the following holds

$$cov(\hat{\mu}_i, y_i) = \sigma^2 \mathbb{E} \left(\frac{\partial \hat{\mu}_i}{\partial y_i} \right).$$

Thus, we have the following unbiased estimator of the degrees of freedom of the fitting procedure $m(\cdot)$

$$\hat{df} = \sum_{i=1}^n \frac{\partial \hat{\mu}_i}{\partial y_i}. \quad (4)$$

Using the degrees of freedom definition as in (3), Efron (2004) employed the covariance penalty approach to prove that the C_p -type statistics (Mallow, 1973) is an unbiased estimator of the true prediction error, where

$$C_p(\hat{\boldsymbol{\mu}}) = \frac{1}{n} \|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2 + \frac{2df(\hat{\boldsymbol{\mu}})}{n} \sigma^2.$$

This reveals the important role played by the degrees of freedom in model assessment. It gives us a principled way of selecting the optimal model without going for computationally expensive methods such as cross-validation, and in certain settings it can offer significantly better prediction accuracy than such methods (Efron, 2004). Many important works followed that of Stein (1981) and Efron (2004). For example, Donoho and Johnstone (1995) used the SURE theory to derive the degrees of freedom for the soft-thresholding operator in wavelet shrinkage; Meyer and Woodroffe (2000) employed this framework to derive the same for shape restricted regression; Li and Zhu (2007) also used this framework to derive an unbiased estimator of the degrees of freedom for penalized quantile regression. Zou et al. (2007) applied the SURE theory for the popular regression shrinkage and variable selection method lasso (Tibshirani, 1996). This is a challenging problem because of the non-linear nature of lasso solution, which does not admit an analytical solution except for certain special

cases. Using sophisticated mathematical analysis, Zou et al. (2007) were able to show that the number of non-zero coefficients provides an unbiased estimate of the degrees of freedom for the lasso. This is a result of great practical importance since this allows one to come up with model selection criteria such as C_p and BIC for the lasso without incurring any extra computational cost.

The degrees of freedom for the reduced rank regression model also proves to be a challenging problem because of the non-linearity of the estimator. As we will see shortly, even though it admits a closed-form solution, the solution is highly non-linear depending on singular value decomposition of certain matrices that in turn depends on the response matrix \mathbf{Y} . In the next section, we study the degrees of freedom of the reduced rank regression model in the framework of SURE and propose a finite-sample exactly unbiased estimator. The importance of such an estimator has been emphasized repeatedly by Shen and Ye (2002), Efron (2004), Zou et al. (2007) among others.

To overcome the analytical difficulty in computing the degrees of freedom, Ye (1998) and Shen and Ye (2002) proposed the generalized degrees of freedom approach, where they evaluate (4) numerically, using data perturbation techniques to compute an approximately unbiased estimator of the degrees of freedom. Efron (2004) also proposed a bootstrap based idea to arrive at an approximately unbiased estimator of (3). Though these kind of simulation based approaches allow us to extend the degrees of freedom approach to many highly non-linear modeling frameworks, they are also computationally expensive.

3 Derivation for the reduced rank regression model

In this section, we propose an analytic method that computes an unbiased estimate of the degrees of freedom for the reduced rank regression model. Recall the multivariate linear regression model as in (1), which can be rewritten as follows

$$\underset{nq \times 1}{vec(\mathbf{Y})} = [\underset{nq \times pq}{I_q \otimes \mathbf{X}}] \underset{pq \times 1}{vec(\mathbf{B})} + \underset{nq \times 1}{vec(\mathbf{E})},$$

where \otimes is the usual Kronecker product between matrices and $vec(\cdot)$ is the column-wise vectorization operator on a matrix. We start by characterizing the solution for the reduced rank regression model. Let $\hat{\mathbf{Y}}_{ols}$ be the usual least squares estimate which admits a singular value decomposition of the form

$$\hat{\mathbf{Y}}_{ols} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{U} \mathbf{D} \mathbf{V}^T, \quad (5)$$

where $\mathbf{U}_{n \times n}$ and $\mathbf{V}_{q \times q}$ are orthogonal matrices that represent the left and right singular vectors and $\mathbf{D}_{n \times q}$ is a diagonal matrix of singular values of $\hat{\mathbf{Y}}_{\text{ols}}$. We will denote the k -th column of \mathbf{U} and \mathbf{V} by u_k and v_k respectively. Using the Eckart-Young theorem (Eckart and Young, 1936), it is not difficult to show that the fitted values for the reduced rank regression model with the estimated coefficients by (2) can be expressed as

$$\hat{\mathbf{Y}}(r) = \hat{\mathbf{Y}}_{\text{ols}} \sum_{k=1}^r v_k v_k^T = \mathbf{U}^r \mathbf{D}^r \mathbf{V}^{rT},$$

where \mathbf{A}^r denotes the first r -columns of a generic matrix \mathbf{A} . Following definition (4), we can naturally define the degrees of freedom for the reduced rank regression model as

$$\hat{df}(r) = \text{tr} \left\{ \frac{\partial \text{vec}(\hat{\mathbf{Y}}(r))}{\partial \text{vec}(\mathbf{Y})} \right\}, \quad (6)$$

where $\text{tr}(\cdot)$ denotes the trace operator for a real square matrix. A direct computation of all partial derivatives within (6), even when being feasible, can be a considerably computational burden when the sample size, n , is large. We reframe the question below so that the number of partial derivatives to be calculated is of order pq and is not a function of n .

Define

$$\mathbf{F} = (\mathbf{X}^T \mathbf{X})^{1/2} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = (\mathbf{X}^T \mathbf{X})^{-1/2} \mathbf{X}^T \mathbf{Y},$$

where $\mathbf{A}^{1/2}$ denotes the square root of a symmetric positive definite matrix \mathbf{A} . Considering the singular value decomposition in (5), it is straightforward to see that \mathbf{F} admits a singular value decomposition of the form

$$\mathbf{F} = \underset{p \times q}{\mathbf{W}} \underset{q \times q}{\mathbf{D}} \underset{q \times q}{\mathbf{V}^T}.$$

Furthermore, with $\hat{\mathbf{Y}}_{\text{ols}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1/2} \mathbf{F}$, we can express the reduced rank regression estimator as,

$$\begin{aligned} \hat{\mathbf{Y}}(r) &= \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1/2} \mathbf{F} \sum_{k=1}^r v_k v_k^T \\ &= \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1/2} \mathbf{W}^r \mathbf{D}^r \mathbf{V}^{rT}. \end{aligned}$$

Now using matrix properties such as $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ and

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C} \otimes \mathbf{A}) \text{vec}(\mathbf{B}), \quad (7)$$

we obtain our unbiased estimator of the degrees of freedom for the reduced rank regression model as,

$$\hat{df}(r) = tr \left\{ \frac{\partial vec(\mathbf{W}^r \mathbf{D}^r \mathbf{V}^{rT})}{\partial vec(\mathbf{F})} \right\}_{pq \times pq}. \quad (8)$$

The details are given in the Appendix.

It is clear, by the chain rule of derivatives, that our remaining task is to compute the following quantities,

$$\left\{ \frac{\partial vec(\mathbf{W})}{\partial vec(\mathbf{F})}, \frac{\partial vec(\mathbf{D})}{\partial vec(\mathbf{F})}, \frac{\partial vec(\mathbf{V})}{\partial vec(\mathbf{F})} \right\}. \quad (9)$$

Note that the singular values and vectors of a matrix are not only highly non-linear functions of the underlying matrix, they are also discontinuous on certain subsets of matrices. This implies that degrees of freedom calculation for the reduced rank regression is a very challenging problem. There has been a considerable amount of work on the smoothness and differentiability of the singular value decomposition of a real matrix in applied mathematics literature. The main result of interest to us states – the singular values and singular vectors of a real matrix \mathbf{F} are smooth differentiable functions of the entries of \mathbf{F} as long as \mathbf{F} has full rank and no repeated singular values. Main references include Magnus and Neudecker (1998), O’Neil (2005) and de Leeuw (2007). In view of this, we will proceed in two main steps to complete the task of obtaining an unbiased estimator of the degrees of freedom for the reduced rank regression model under the SURE framework:

1. Derive the partial derivatives in (9) over the set of matrices as described above.
2. Prove that the set where the partial derivatives do not exist has Lebesgue measure 0.

The following two subsections will address the aforementioned steps respectively and thus will complete the derivation.

3.1 Jacobian of the singular value decomposition

In this subsection, we propose a method that computes the partial derivatives of the singular values and singular vectors of a matrix with respect to the entries of the underlying matrix, and we acknowledge that we borrow the idea from Wright (1992) and Papadopoulos and Lourakis (2000). We will present the results in the context of a generic $p \times q$ matrix \mathbf{F} , and we will only present the case where $p \geq q$; the reverse scenario follows similarly. Assume that \mathbf{F} admits a singular value decomposition,

$$\mathbf{F} = \underset{p \times p}{\mathbf{W}} \underset{p \times q}{\mathbf{D}} \underset{q \times q}{\mathbf{V}^T},$$

such that $\mathbf{W}^T \mathbf{W} = \mathbf{W} \mathbf{W}^T = \mathbf{I}_p$ and $\mathbf{V}^T \mathbf{V} = \mathbf{V} \mathbf{V}^T = \mathbf{I}_q$. Define $d_k = d_{kk}$ to be the k -th diagonal element of \mathbf{D} ; $\mathbf{\Omega}_W^{jk} = \mathbf{W}^T \partial \mathbf{W} / \partial f_{jk}$, where f_{jk} is the (j, k) th entry of \mathbf{F} , and equivalently define $\mathbf{\Lambda}_V^{jk} = \left(\mathbf{\Omega}_{V^T}^{jk} \right)^T = \partial \mathbf{V}^T / \partial f_{jk} \mathbf{V}$. Then, by noting $\mathbf{W}^T \mathbf{W} = \mathbf{I}_p$ and $\mathbf{V} \mathbf{V}^T = \mathbf{I}_q$, we obtain $\partial \mathbf{W} / \partial f_{jk} = \mathbf{W} \mathbf{\Omega}_W^{jk}$ and $\partial \mathbf{V} / \partial f_{jk} = -\mathbf{V} \mathbf{\Lambda}_V^{jk}$. Consequently, our task is equivalent to obtaining $\mathbf{\Omega}_W^{jk}$, $\mathbf{\Lambda}_V^{jk}$ and $\partial d_k / \partial f_{jk}$, for $j = 1, 2, \dots, p$; $k = 1, 2, \dots, q$, which can be completed by

(i) Showing

$$\frac{\partial d_k}{\partial f_{jk}} = w_{jk} v_{kk}, \quad k = 1, 2, \dots, q. \quad (10)$$

(ii) Obtaining $\mathbf{\Omega}_W^{jk}(j', k')$ and $\mathbf{\Lambda}_V^{jk}(j', k')$ for $j' \neq k', 1 \leq j', k' \leq q$ by solving a set of systems of linear equations,

$$\begin{cases} d_{k'} \mathbf{\Omega}_W^{jk}(j', k') + d_{j'} \mathbf{\Lambda}_V^{jk}(j', k') = w_{jj'} v_{kk'} \\ d_{j'} \mathbf{\Omega}_W^{jk}(j', k') + d_{k'} \mathbf{\Lambda}_V^{jk}(j', k') = -w_{jk'} v_{kj'} \end{cases} \quad \text{for } j' = 1, 2, \dots, q-1; k' = j' + 1, \dots, q. \quad (11)$$

(iii) Obtaining $\mathbf{\Omega}_W^{jk}(j', k')$ for $q+1 \leq j' \leq p$ and $1 \leq k' \leq q$ by solving

$$d_{j'} \mathbf{\Omega}_W^{jk}(j', k') = -w_{jk'} v_{kj'} \quad \text{for } j' = q+1, \dots, p; k' = 1, \dots, q. \quad (12)$$

Note that we can solve for $\mathbf{\Lambda}_V^{jk}$ fully, while for $\mathbf{\Omega}_W^{jk}$ we are only able to get a partial solution, where the last $p \times (p-q)$ block remains undetermined. Even though the partial solution of $\mathbf{\Omega}_W^{jk}$ implies that we can only solve for the first q columns of $\partial \mathbf{W} / \partial f_{jk}$, it is sufficient since the first q columns of \mathbf{W} are enough to determine the singular value decomposition of an $p \times q$ matrix with $p \geq q$.

We complete this subsection by noting that the items (i)–(iii) above are direct consequences of pre-multiplying by \mathbf{W}^T and post-multiplying by \mathbf{V} the following equation

$$\frac{\partial \mathbf{F}}{\partial f_{jk}} = \frac{\partial \mathbf{W}}{\partial f_{jk}} \mathbf{D} \mathbf{V}^T + \mathbf{W} \frac{\partial \mathbf{D}}{\partial f_{jk}} \mathbf{V}^T + \mathbf{W} \mathbf{D} \frac{\partial \mathbf{V}^T}{\partial f_{jk}},$$

which leads to

$$\mathbf{W}^T \frac{\partial \mathbf{F}}{\partial f_{jk}} \mathbf{V} = \mathbf{\Omega}_W^{jk} \mathbf{D} + \frac{\partial \mathbf{D}}{\partial f_{jk}} + \mathbf{D} \mathbf{\Lambda}_V^{jk}. \quad (13)$$

Item (i) and (10) are the result of matching the diagonal entries in (13) and using the fact that $\mathbf{\Omega}_W^{jk}$ and $\mathbf{\Lambda}_V^{jk}$ are anti-symmetric matrices; a matrix A is anti-symmetric if $A + A^T = 0$. Note that both sides of (13) are $p \times q$ matrices with $p \geq q$. Items (ii) and (iii) are obtained by matching the upper $q \times q$ submatrices and by matching the lower $(p-q) \times q$ submatrices from the two sides of (13), respectively.

3.2 Existence of partial derivatives almost everywhere

Studying the system of linear equations (11) and (12), it is not difficult to find that the solution exists only if,

$$d_1 > d_2 > \dots > d_q > 0.$$

That is, $\hat{\mathbf{Y}}_{\text{OLS}}$ must be of rank q and it cannot have repeated singular values as described in O'Neil (2005). The following theorem gives the main result that allows us to apply Stein's framework in deriving the degrees of freedom of the reduced rank regression problem.

Theorem 3.1. *Let $\mathbb{R}^{p \times q}$ be the space of all real-valued $p \times q$ dimensional matrices equipped with the Lebesgue measure μ . Also, let $\mathcal{S} \subseteq \mathbb{R}^{p \times q}$ denote the subset of matrices that have full rank and no repeated singular values. Then*

$$\mu(\mathcal{S}) = 1.$$

To prove the theorem, we start with a few definitions and facts from algebraic geometry and matrix analysis.

Definition 3.2. An algebraic variety over \mathbb{R}^k (or \mathbb{C}^k) is defined as the set of points satisfying a system of polynomial equations $\{f_\ell(x_1, x_2, \dots, x_k) = 0; \ell \in \mathcal{I}\}$.

Here each $f_\ell(\cdot)$ is a polynomial function of its arguments and \mathcal{I} denotes an index set. If at least one of the $f_\ell(\cdot) \not\equiv 0$, then it is called a *proper subvariety*. Note that a proper subvariety must be of dimension less than k and therefore has Lebesgue measure 0 in \mathbb{R}^k (Allman et al., 2009). For a more detailed discussion, we recommend Hartshorne (1977) or Cox et al. (2007).

Proposition 3.3. (Laub, 2004) *Any square symmetric matrix $\mathbf{M} \in \mathbb{R}^{k \times k}$ has at least one repeated eigenvalue if and only if $\text{rank}(\mathbf{M} \otimes \mathbf{I}_k - \mathbf{I}_k \otimes \mathbf{M}) < (k^2 - k)$.*

Now we prove the theorem. First we define

$$\begin{aligned} \mathcal{S}_1 &= \{\mathbf{A} \in \mathbb{R}_{p \times q} : \mathbf{A} \text{ has at least one 0 singular value}\}, \\ \mathcal{S}_2 &= \{\mathbf{A} \in \mathbb{R}_{p \times q} : \mathbf{A} \text{ has at least one repeated singular value}\}. \end{aligned}$$

Note that $\mathcal{S}^c = \mathcal{S}_1 \cup \mathcal{S}_2$, thus it is enough to show that $\mu(\mathcal{S}_1) = 0$ and $\mu(\mathcal{S}_2) = 0$. By definition 3.2 and the discussion above it suffices to show that \mathcal{S}_1 and \mathcal{S}_2 are proper subvarieties of $\mathbb{R}^{p \times q}$. \mathcal{S}_1 can be rewritten as follows

$$\mathcal{S}_1 = \{\mathbf{A} \in \mathbb{R}_{p \times q} : \det(\mathbf{A}^T \mathbf{A}) = 0\}.$$

Here $\det(\cdot)$ denotes the determinant operator for a square matrix. Note that $\det(\mathbf{A}^T \mathbf{A})$ is a non-trivial polynomial in entries of \mathbf{A} and hence \mathcal{S}_1 is a proper subvariety and has Lebesgue

measure 0. For \mathcal{S}_2 note that if $\mathbf{A} \in \mathbb{R}^{p \times q}$ has at least one repeated singular value, it implies that $\mathbf{A}^T \mathbf{A} \in \mathbb{R}^{p \times p}$ has at least one repeated eigenvalue. Then in view of proposition 3.3, \mathcal{S}_2 can be reformulated as

$$\mathcal{S}_2 = \{ \mathbf{A} \in \mathbb{R}_{p \times q} : \text{rank}(\mathbf{A}^T \mathbf{A} \otimes \mathbf{I}_q - \mathbf{I}_q \otimes \mathbf{A}^T \mathbf{A}) < (q^2 - q) \}.$$

This is an algebraic variety since it can be expressed as the solution to all minors of order $\geq (q^2 - q)$ being equal to 0, which are all polynomial equations in the entries of \mathbf{A} . Thus, we have shown that, $\mu(\mathcal{S}_1 \cup \mathcal{S}_2) = 0$.

3.3 Computational cost

It is fairly straightforward to derive the computational complexity of the proposed method. For each $j = 1, 2, \dots, p$ and $k = 1, 2, \dots, q$, we need to solve pq many 2×2 systems of linear equations. Hence the complexity of the procedure is of $\mathcal{O}(p^2 q^2)$ once the initial singular value decomposition has been carried out. On the other hand, for the data perturbation method (Ye, 1998), we need to compute a singular value decomposition for each $i = 1, 2, \dots, n$ and $k = 1, 2, \dots, q$ to numerically evaluate the derivatives. The singular value decomposition of a $n \times q$ matrix has a computational complexity of $\mathcal{O}(nq \max\{n, q\})$. Therefore the data perturbation approach would lead to a complexity of $\mathcal{O}(n^2 q^2 \max\{n, q\})$, which is much higher than that of the exact method we propose. In addition, such data perturbation methods can often lead to “unstable” estimates. This is because even the state-of-art algorithms for computing the singular value decomposition, e.g. LAPACK (Anderson et al., 1994) is not very robust with respect to small perturbations. The computational cost of the bootstrap method (Efron, 2004) is $\mathcal{O}(Bnq \max\{n, q\})$, which in general is also inferior to $\mathcal{O}(p^2 q^2)$.

4 Extension to other reduced rank methods

This section deals with the extension of the proposed degrees of freedom technique to two other recently proposed rank regularized multivariate regression methods (Mukherjee and Zhu, 2011; Chen et al., 2012b). This shows that the technique developed in the previous section is quite general and can be applied to several other problems that deal with rank regularization or singular value thresholding. This allows us to do a principled model selection and avoid the computationally expensive cross-validation method for a wider range of problems.

4.1 Extension to reduced rank ridge regression

Mukherjee and Zhu (2011) proposed a modified version of reduced rank regression model, which incorporates an additional ridge penalty to counter the issue of collinearity among the prediction variables. In particular, they choose to minimize

$$\hat{\mathbf{B}}(\lambda, r) = \arg \min_{\{\mathbf{B}: \text{rank}(\mathbf{B}) \leq r\}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda \|\mathbf{B}\|_F^2, \quad (14)$$

where $1 \leq r \leq \min\{p, q\}$ and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. The additional ridge penalty is motivated by the fact that it ensures a well-behaved estimator of \mathbf{B} even in the presence of high-collinearity, whereas the rank penalty encourages dimension reduction. Note that this method has two tuning parameters, (λ, r) . The authors propose a k -fold cross-validation approach to select the optimal model over a grid of tuning parameters. The degrees of freedom methodology developed in section 3 can be easily extended here. We start by describing the solution to the optimization problem (14). Denote the usual ridge regression estimator by

$$\hat{\mathbf{Y}}_R = \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y} = \tilde{\mathbf{U}} \tilde{\mathbf{D}} \tilde{\mathbf{V}}^T,$$

where the rightmost part denotes the singular value decomposition. Following similar steps as in the case of reduced rank regression, we can show that the solution to (14) is given by

$$\hat{\mathbf{Y}}(\lambda, r) = \tilde{\mathbf{U}}^r \tilde{\mathbf{D}}^r \tilde{\mathbf{V}}^{rT}.$$

Now, using the same chain of arguments as before it is straightforward to deduce that the degrees of freedom estimator for the reduced rank ridge regression model under Stein's framework is

$$\begin{aligned} \hat{df}(\lambda, r) &= \text{tr} \left(\frac{\partial \text{vec}(\hat{\mathbf{Y}}(\lambda, r))}{\partial \text{vec}(\hat{\mathbf{Y}}_R)} \frac{\partial \text{vec}(\hat{\mathbf{Y}}_R)}{\partial \text{vec}(\mathbf{Y})} \right) \\ &= \text{tr} \left(\frac{\partial \text{vec}(\tilde{\mathbf{U}}^r \tilde{\mathbf{D}}^r \tilde{\mathbf{V}}^{rT})}{\partial \text{vec}(\hat{\mathbf{Y}}_R)} [\mathbf{I}_q \otimes \mathbf{X}(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T] \right). \end{aligned} \quad (15)$$

To compute the first part of (15), i.e., the partial derivatives of singular values and vectors of a matrix with respect to the entries of the underlying matrix, we can employ the technique developed in subsection 3.1. Once (15) is computed, we can use any of the popular model selection criteria such as Mallows's C_p , BIC or GCV to select an optimal model over a grid of tuning parameter values.

4.2 Extension to the adaptive SVD soft-thresholding estimator

Recently, Chen et al. (2012b) proposed a very interesting method that they call adaptive SVD soft-thresholding estimator for multivariate regression. The rank penalty that we have discussed in great detail can be considered as an ℓ_0 penalty on the singular values of \mathbf{B} . Yuan et al. (2007) proposed a factor estimation and selection procedure, which penalizes the ℓ_1 norm of the singular values of \mathbf{B} (nuclear norm) to encourage sparsity. But the resulting optimization problem is rather challenging and does not admit any closed form solution. Chen et al. (2012b) choose to introduce a weighted nuclear norm penalty on the signal matrix \mathbf{XB} . Specifically, the resulting optimization problem is

$$\hat{\mathbf{B}}(\lambda) = \arg \min_{\mathbf{B}} \frac{1}{2} \|\mathbf{Y} - \mathbf{XB}\|_F^2 + \lambda \sum_{k=1}^{\min\{p,q\}} w_k d_k(\mathbf{XB}), \quad (16)$$

where $d_k(\mathbf{A})$ denotes the k -th largest singular value of a generic matrix \mathbf{A} . The weights w_k are assumed to be non-negative and non-decreasing to ensure more severe penalization for smaller singular values. The advantages of this choice of penalty function are two-fold. Firstly, it focuses directly on the singular values of the signal matrix \mathbf{XB} instead of the coefficient matrix \mathbf{B} , which should lead to better prediction performance especially when \mathbf{X} is highly collinear. Secondly, this leads to a closed form solution to (16). Before getting into the issue of model complexity, we first describe the solution to the optimization problem (16). We will continue to use the notation for the least squares estimator $\hat{\mathbf{Y}}_{\text{ols}}$ described in (5). Define the adaptive soft-thresholding operator for a diagonal matrix $\mathbf{A}_{m \times m}$ as

$$\mathcal{S}_{\lambda w}(\mathbf{A}) = \text{Diag}\{(A_k - \lambda w_k)_+ : k = 1, 2, \dots, m\},$$

where $(\cdot)_+$ denotes the positive part. The authors show that the estimated response matrix for the proposed method is given by

$$\hat{\mathbf{Y}}(\lambda) = \mathbf{U} \mathcal{S}_{\lambda w}(\mathbf{D}) \mathbf{V}^T.$$

Chen et al. (2012b) utilizes this closed form solution to prove desirable theoretical properties such as rank consistency and error bounds. But for practical purposes we still need a way of efficiently choosing the tuning parameter λ . The authors propose two ways of doing that: the first one is based on computationally intensive cross-validation, and the second one is generalized cross-validation (GCV). For the GCV approach they choose to use the number of free parameters heuristic (Davies and Tso, 1982; Reinsel and Velu, 1998).

It turns out that the degrees of freedom technique for rank regularization can also be extended to this type of weighted ℓ_1 penalty on singular values. Specifically, as done in (15), we have

$$\begin{aligned}\widehat{df}(\lambda) &= \text{tr} \left(\frac{\partial \text{vec}(\widehat{\mathbf{Y}}(\lambda))}{\partial \text{vec}(\widehat{\mathbf{Y}}_{\text{ols}})} \frac{\partial \text{vec}(\widehat{\mathbf{Y}}_{\text{ols}})}{\partial \text{vec}(\mathbf{Y})} \right) \\ &= \text{tr} \left(\frac{\partial \text{vec}(\mathbf{U}\mathbf{S}_{\lambda w}(\mathbf{D})\mathbf{V}^T)}{\partial \text{vec}(\widehat{\mathbf{Y}}_{\text{ols}})} [\mathbf{I}_q \otimes \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] \right).\end{aligned}\quad (17)$$

Once again, it is important to note that in order to obtain the estimator (17), it suffices to compute the partial derivatives of the singular values and singular vectors with respect to the underlying matrix, $\widehat{\mathbf{Y}}_{\text{ols}}$, which is what we are able to compute using the techniques developed in subsection 3.1.

5 Simulation studies

In this section, we evaluate the performance of the proposed method by simulation studies. Specifically, we aim to demonstrate two things: 1) the exact unbiased estimator of the degrees of freedom for the reduced rank regression is in general significantly higher than the naive estimator; 2) using the exact estimator of the degrees of freedom enables us to gain prediction accuracy over the naive estimator.

5.1 Unbiasedness

In this simulation, we aim to show that the degrees of freedom estimator defined via (8) is unbiased and it can be significantly higher than the naive estimator that simply counts the number of free parameters. Here unbiasedness is defined over the error distribution, and we treat \mathbf{X} as a fixed design matrix. Parameters of the study are as follows, $n = 200$, $p = 15$ and $q = 12$. Let Σ denote the covariance matrix of the predictor variables, \mathbf{X} . We consider a low level of correlation among \mathbf{X} , namely $\Sigma_{jj'} = 0.3^{|j-j'|}$. Rows of the predictor matrix are generated independently from $N_p(\mathbf{0}, \Sigma)$. To control the singular structure of \mathbf{B} through the covariance of signals \mathbf{XB} , $\mathbf{B}^T\Sigma\mathbf{B}$, we take the left singular vectors of \mathbf{B} the same as the eigenvectors of Σ , whereas the right singular vectors of \mathbf{B} are generated by orthogonalizing a random standard normal matrix. The difference between successive non-zero singular value of \mathbf{B} is fixed at 2. The error matrix is generated from i.i.d. standard normal distribution. We replicate the process 500 times; note that the design matrix remains fixed. We compare the proposed exact method against the data perturbation technique (Ye, 1998) and the parametric bootstrap (Efron, 2004). For the data perturbation method, we consider 25 perturbations

of the response matrix for each replication to estimate the partial derivatives numerically. We used the choice of 0.1σ for the perturbation size, where σ is the error standard deviation. For the parametric bootstrap approach, we choose $B = 50$ bootstrap estimates for each replication to compute the covariance estimator using (3). The mean of the bootstrapped covariance estimator over the replications can be thought of as a Monte-Carlo estimator of the true degrees of freedom. This will be denoted by $df(r)$. Ideally we would expect the proposed exact estimator to be fairly close to $df(r)$ on average. We compare estimators against the naive degrees of freedom estimate namely, $df_n(r) = r(p + q - r)$, which denotes the number of free parameters in a $p \times q$ matrix of rank r . Note that the naive estimator does not depend on the data.

First we consider the situation where the true underlying \mathbf{B} has full rank. On the left of Figure 1 we see that out of the four estimators, i.e. the proposed exact method, the data perturbation estimator, the parametric bootstrap estimator and the naive estimator, the averages of the first three are almost indistinguishable to the naked eye and they are significantly higher (often 20% or 30% higher) than the of the naive estimator, which counts the number of free parameters. The figure on the right side allows us to get a sense of the variability of the estimation procedures. Clearly the standard error for the exact method is orders of magnitudes smaller than that of either data perturbation or parametric bootstrap estimators.

Next we change the rank of \mathbf{B} to 8 and repeat the same experiment. Once again we find in the panel on the left side of Figure 2 that the averages resulted by the first three estimation methods nearly overlap, and they are consistently higher than the naive estimate. The figure also reveals an interesting pattern, that is, the exact estimators seem to match the naive estimator very closely at the true rank, i.e. 8 in this case. We could not explain this behaviour theoretically at this point but this was observed in a wide variety of simulation studies. In addition, we find in the figure on the right that the standard error of the exact degrees of freedom becomes drastically higher once we go above the true underlying rank. This arises from the fact that once we go above the true rank, the singular values of $\hat{\mathbf{Y}}_{\text{ols}}$ basically correspond to noise, and can be very close to each other. Hence slight perturbations of the data might lead to different singular directions being selected, which implies higher variability in model complexity. This has also been noted by Ye (1998), that is, if we are trying to fit pure error components, the degrees of freedom tends to be higher and unstable.

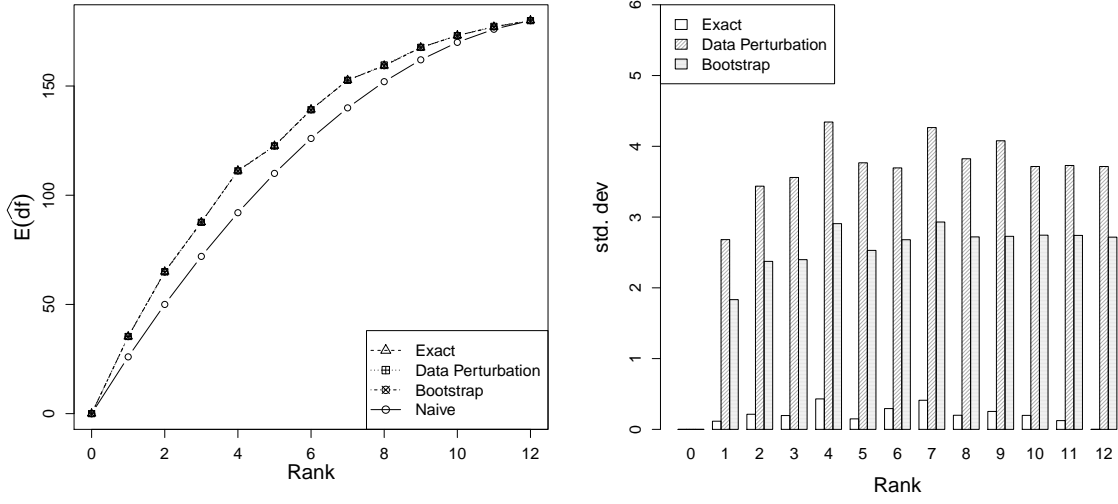


Figure 1: \mathbf{B} is full rank. Left: the average degrees of freedom estimate over 500 replications; Right: the standard error of the degrees of freedom estimate over 500 replications.

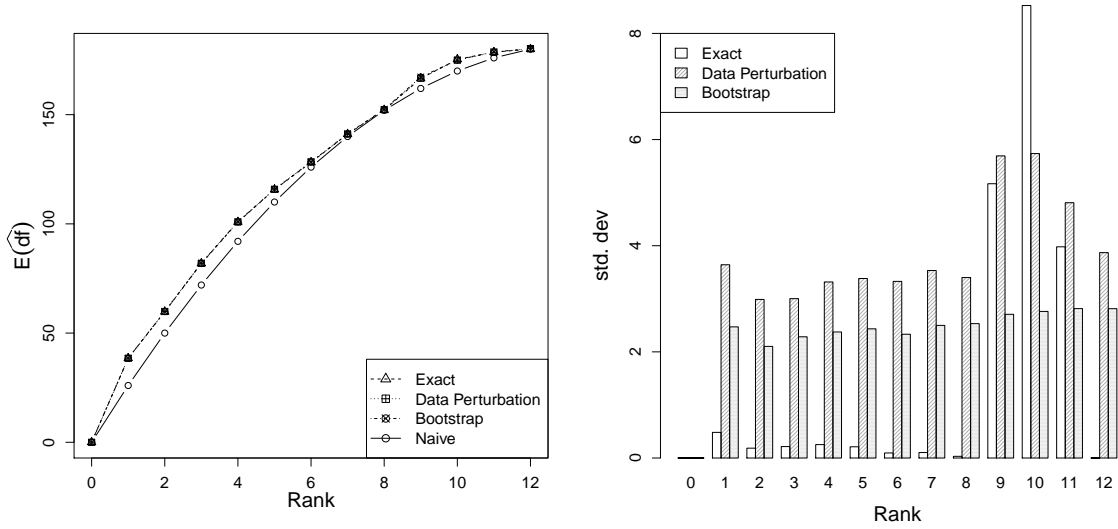


Figure 2: \mathbf{B} has rank 8. Left: average degrees of freedom estimate over 500 replications; Right: the standard error of the degrees of freedom estimate over 500 replications.

5.2 Prediction performance

The previous set of simulations have shown that the exact degrees of freedom estimator can be significantly different from the number of free parameters estimator. Degrees of freedom estimates are commonly used in various model selection criteria. In this subsection, we aim to show that for reduced rank regression, we can gain in prediction accuracy by using the exact degrees of freedom estimator in a model selection criterion instead of the naive estimator. Since our focus is on prediction accuracy, we consider Mallow's C_p (Mallow, 1973) and generalized cross-validation (GCV) (Golub et al., 1979) as our model selection criteria. GCV has the added advantage that it does not require an estimate for the error variance and hence proven to be much popular in practice. In the context of reduced rank regression, they are defined as follows

$$\begin{aligned} C_p(r) &= \frac{1}{nq} \left(\|\mathbf{Y} - \hat{\mathbf{Y}}(r)\|_F^2 + 2\hat{\sigma}^2 df(r) \right), \\ GCV(r) &= \frac{nq \|\mathbf{Y} - \hat{\mathbf{Y}}(r)\|_F^2}{(nq - df(r))^2}. \end{aligned}$$

We select the model that minimizes the respective criterion over $1 \leq r \leq \min\{p, q\}$. For this study, we choose $p = 12$, $q = 10$ and fix the true rank of \mathbf{B} at 3. We consider two different sample sizes $n = 50$ and 200 as well as two levels for error variance, namely, $\sigma^2 = 1$ and 2 . Correlation among predictor variables is kept at a moderate level of 0.5 . We consider two specific structures for the singular values of \mathbf{B}

$$\begin{aligned} \text{Model I} &: \sigma(\mathbf{B}) = \{2.5, 1.0, 0.4, 0, \dots, 0\} \Rightarrow \sigma(\mathbf{B}^T \Sigma \mathbf{B}) = \{17.23, 6.21, 0.26, 0, \dots, 0\}, \\ \text{Model II} &: \sigma(\mathbf{B}) = \{2.5, 2.0, 1.5, 0, \dots, 0\} \Rightarrow \sigma(\mathbf{B}^T \Sigma \mathbf{B}) = \{17.23, 8.82, 3.70, 0, \dots, 0\}. \end{aligned}$$

From Model I to Model II, the non-zero singular values of \mathbf{B} increase, which in turn implies that the non-zero singular values of $\mathbf{B}^T \Sigma \mathbf{B}$ increase, thus making the problem easier both in terms of signal to noise ratio (SNR) and the gap between the smallest non-zero singular value and 0. The latter was shown to play an important role in rank recovery and prediction performance by Bunea et al. (2011). Data is generated using the same procedure as in the previous study except that \mathbf{X} is no longer fixed. We compare the prediction performance between several model selection criteria by the model error, which is defined as

$$\text{ME} = \text{tr} \left\{ (\mathbf{B} - \hat{\mathbf{B}})^T \Sigma (\mathbf{B} - \hat{\mathbf{B}}) \right\}.$$

We repeat the simulation for 100 times at each combination of model, sample size and error variance. Recall that we will be comparing four model selection criteria, namely, C_p with the exact degrees of freedom ($C_p(\text{e})$), C_p with the naive degrees of freedom ($C_p(\text{n})$), GCV

with the exact degrees of freedom (GCV(e)) and GCV with the naive degrees of freedom (GCV(n)). We fit the optimal model based on each of the competing model selection criteria and compute the model error. Table 1 summarizes the results.

Table 1: Prediction accuracy comparison of different model selection criteria, namely, $C_p(e)$, $C_p(n)$, GCV(e), GCV(n)

Sample Size			$n = 50$			
Error Var	Model		$C_p(e)$	$C_p(n)$	GCV(e)	GCV(n)
$\sigma^2 = 1$	I		1.408(0.31)	1.543(0.39)	1.393(0.31)	1.523(0.39)
	II		1.541(0.41)	1.636(0.56)	1.541(0.41)	1.627(0.54)
$\sigma^2 = 2$	I		2.643(0.70)	2.857(0.83)	2.631(0.69)	2.831(0.83)
	II		3.231(0.89)	3.421(1.17)	3.231(0.89)	3.386(1.09)
Sample Size			$n = 200$			
Error Var	Model		$C_p(e)$	$C_p(n)$	GCV(e)	GCV(n)
$\sigma^2 = 1$	I		0.344(0.10)	0.361(0.11)	0.344(0.10)	0.362(0.11)
	II		0.310(0.07)	0.343(0.10)	0.310(0.07)	0.343(0.10)
$\sigma^2 = 2$	I		0.708(0.16)	0.724(0.21)	0.708(0.16)	0.724(0.21)
	II		0.605(0.12)	0.650(0.17)	0.605(0.12)	0.650(0.17)

We find that for both C_p and GCV, using the proposed exact degrees of freedom estimator performs better in terms of prediction accuracy than its naive counterpart. It has lower average prediction error as well as smaller standard error in all of the combinations of model, sample size and error variance. The relative gain is larger when the sample size is small and the smallest non-zero singular value of the limiting signal matrix is closer to 0. To better understand the advantage of using the exact degrees of freedom in a model selection criterion, we have also summarized in Table 2 the frequency of the rank selection among the 100 replications at each combination of model and sample size when $\sigma^2 = 1$. We have skipped the other case ($\sigma^2 = 2$) for brevity.

First focus on the combination of Model I and small sample size, i.e. $n = 50$. Recall that the smallest non-zero singular value of $\mathbf{B}^T \Sigma \mathbf{B}$ for Model I is rather small, i.e. 0.26, implying the effective rank for prediction purpose is 2 instead of 3. Interestingly we find that the C_p and GCV criteria with the exact estimator for the degrees of freedom select the rank 2 model nearly 90% of times. On the other hand, the same criteria equipped with the naive degrees

Table 2: Comparison of rank selection performance of the competing model selection criteria, $C_p(e)$, $C_p(n)$, GCV(e) and GCV(n).

Sample Size	Rank	Model I			
		$C_p(e)$	$C_p(n)$	GCV(e)	GCV(n)
$n = 50$	2	88	62	91	65
	3	12	37	9	35
	4	0	1	0	0
	5	0	0	0	0
$n = 200$	2	4	1	4	1
	3	89	80	89	79
	4	5	17	5	18
	5	2	2	2	2
Sample Size	Rank	Model II			
		$C_p(e)$	$C_p(n)$	GCV(e)	GCV(n)
$n = 200$	2	0	0	0	0
	3	95	85	95	85
	4	5	14	5	15
	5	0	1	0	0
$n = 200$	2	0	0	0	0
	3	98	81	98	81
	4	2	17	2	17
	5	0	2	0	2

of freedom estimator has a more even distribution, approximately 65% - 35% of selecting the rank 2 versus the rank 3 model, but it performs sub-optimally to the exact estimator in terms of prediction accuracy. This is due to the bias-variance trade-off in the mean squared error. The third largest singular direction contains very little explanatory power, thus a simpler model is able to attain better prediction accuracy especially for small sample sizes. When we move to higher sample size of $n = 200$, we find that the rank selection performances of both approaches improve but the proposed method still does better than the naive approach in terms of rank selection. In Model II, the signal strength and the smallest non-zero singular value are both higher, leading to a much better rank selection performance even at the smaller sample size, but once again, the model selection criteria with the exact degrees of freedom do better. Overall it appears that the naive estimator tends to overestimate the

rank slightly, which is expected since it underestimates the degrees of freedom of the reduced rank regression model.

6 Analysis of arabidopsis thaliana data

In this section, we apply the proposed method to a genetic association data set that was published in Wille et al. (2004). This is a microarray experiment aimed at understanding the regulatory control mechanisms between the isoprenoid gene network in *Arabidopsis thaliana* plant (more commonly known as thale cress or mouse-ear cress). It is known that isoprenoids serve many important biochemical functions in plants. To monitor the gene-expression levels, 118 GeneChip microarray experiments were carried out. The predictors consist of 39 genes from two isoprenoid bio-synthesis pathways namely MVA and MEP, whereas the responses consist of gene-expression of 795 genes from 56 metabolic pathways, many of which are downstream of the two pathways considered as predictors. Thus some of the responses are expected to show significant associations to the predictor genes. To facilitate it further, we select two downstream pathways namely, Carotenoid and Phytosterol as our responses. It has already been proven experimentally that the Carotenoid pathway is strongly attached to the MEP pathway, whereas the Phytosterol pathway is significantly related to the MVA pathway. See Wille et al. (2004) and the references therein for a more detailed discussion on the biological aspects. Finally we have 118 observations on $p = 39$ predictors and $q = 36$ responses. All the predictors and responses are log-transformed to reduce the skewness of the data. We also standardize the responses in order to make them comparable.

We split the data set randomly into training and test sets of equal size. The model is fit using the training samples and then we use it to predict on the test set. The performance measure under consideration is the usual mean squared prediction error

$$\text{MSPE} = \frac{2}{nq} \|\mathbf{Y}_{\text{test}} - \hat{\mathbf{Y}}_{\text{test}}\|_F^2. \quad (18)$$

Here $\hat{\mathbf{Y}}$ denotes the predicted value and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. The entire process is repeated 100 times based on random splits to ensure that the results remain robust to the process of splitting. Model selection methods under consideration are Mallows's C_p , GCV and BIC with the exact degrees of freedom as in (8) and the naive degrees of freedom which counts the number of free parameters in a matrix of restricted rank.

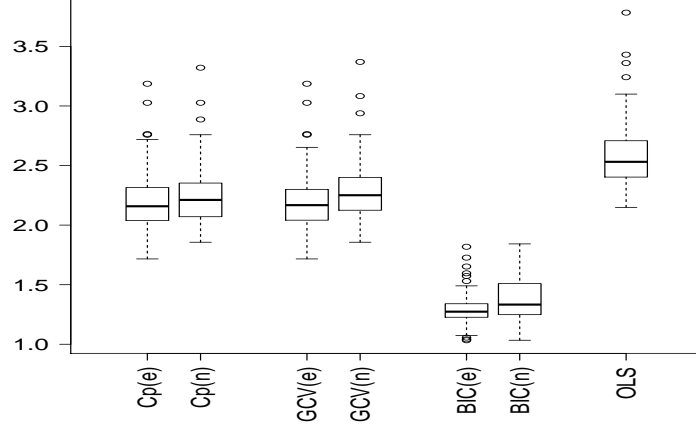


Figure 3: Boxplot of mean square prediction error of each method over 100 random splits.

The mean squared prediction errors for each method are summarized using the boxplot in Figure 3. As we can see, for all three model selection criteria considered, the use of the exact unbiased estimator enables us to outperform the one which uses the heuristic estimator in terms of prediction accuracy.

Table 3: Prediction accuracy and rank selection performance for the competing methods on the *Arabidopsis thaliana* data.

	Cp(e)	Cp(n)	GCV(e)	GCV(n)	BIC(e)	BIC(n)	OLS
Avg(Pred Err)	2.197	2.243	2.192	2.282	1.297	1.387	2.589
Std(Pred Err)	0.250	0.246	0.248	0.246	0.134	0.201	0.282
Mean(Est Rank)	8.760	9.710	8.680	10.520	1.090	1.480	–

Table 3 shows the average prediction error over the 100 random splits as well as its standard deviation. This also reflects that the model selection criteria with the exact unbiased estimate of the degrees of freedom tends to select smaller rank solutions than that corresponding to the heuristic estimate of number of free parameters. This is to be expected since the exact unbiased estimate is usually higher than the number of free parameters and thus penalizes more severely on models with higher complexity.

7 Concluding remarks

We have proposed an exact unbiased estimator of the degrees of freedom for the reduced rank regression model in the framework of SURE. Through simulations and an Arabidop-

sis thaliana data example, we show that the exact unbiased estimator can be significantly different than the heuristic estimator that simply counts the number of free parameters. The proposed estimator matches with the numerical estimators due to Ye (1998) and Efron (2004), which further verifies the validity of our method. It gives us a principled way of using various model selection criteria, such as, C_p , GCV and BIC to select an optimal rank solution. The use of the proposed estimator often outperforms that of the heuristic counterpart. The new degrees of freedom estimator can be computed easily by solving a set of systems of linear equations without incurring much computational cost. This feature and the potentially inflated variation are the main drawback for data perturbation or parametric bootstrap estimators. The new estimator improves both aspects. We have also shown that the methods developed here are quite general and can be extended to other related estimation procedures that employ regularization of the singular values, such as the reduced rank ridge regression (Mukherjee and Zhu, 2011) and a weighted nuclear norm penalized reduced rank regression (Chen et al., 2012b). The nuclear norm regularized multivariate regression has attracted much attention in recent years, e.g. Yuan et al. (2007), Bunea et al. (2011) and Chen et al. (2012b). The resulting optimization problems are often solved via iterative algorithms and do not have closed form solutions except for certain special cases such as the orthogonal design matrix \mathbf{X} . It is our future plan to extend the proposed degrees of freedom approach to the nuclear norm penalization, which is certainly a challenging problem.

References

- Allman, E., Matias, C., and Rhodes, J. (2009), “Identifiability of parameters in latent structure models with many observed variables,” *Annals of Statistics*, 37, 3099–3132.
- Anderson, E., Bai, Z., Bischof, C., Demmel, J., Dongarra, J., Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Ostrouchov, S., and Sorensen, D. (1994), *LAPACK Users’ Guide*, Society for Industrial and Applied Mathematics.
- Anderson, T. (1951), “Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions,” *Annals of Mathematical Statistics*, 22, 327–351.
- (1999), “Asymptotic Distribution of the Reduced Rank Regression Estimator under General Conditions,” *Annals of Statistics*, 27, 1141–1154.
- (2002), “Specification and Misspecification in Reduced Rank Regression,” *Sankhyā, Series A*, 64, 193–205.

- Breiman, L. and Friedman, J. (1997), “Predicting multivariate responses in multiple linear regression,” *Journal of the Royal Statistical Society, Series B*, 59, 3–37.
- Bunea, F., She, Y., and Wegkamp, M. (2011), “Optimal Selection of Reduced Rank Estimators of High-Dimensional Matrices,” *Annals of Statistics*, 39, 1282–1309.
- Chen, K., Chan, K., and Stenseth, N. (2012a), “Reduced-rank stochastic regression with a sparse singular value decomposition,” *Journal of the Royal Statistical Society, Series B*, 74, 203–221.
- Chen, K., Dong, H., and Chan, K. (2012b), “Adaptive SVD Soft-Thresholding Estimators in Multivariate Regression,” <http://arxiv.org/pdf/1201.0381.pdf>.
- Cox, D., Little, J., and O’Shea, D. (2007), *Ideals, Varieties, and Algorithms. An introduction to computational algebraic geometry and commutative algebra*, Springer, New York, 3rd ed.
- Davies, P. and Tso, M. (1982), “Procedures for Reduced-Rank Regression,” *Applied Statistics*, 31, 244–255.
- de Leeuw, J. (2007), “Derivatives of Generalized Eigen Systems with Applications,” *Department of Statistics, UCLA*, 1–28.
- Donoho, D. and Johnstone, I. (1995), “Adapting to Unknown Smoothness via Wavelet Shrinkage,” *Journal of the American Statistical Association*, 90, 1200–1224.
- Eckart, C. and Young, G. (1936), “The Approximation of One Matrix by Another of Lower Rank,” *Psychometrika*, 1, 211–218.
- Efron, B. (2004), “The Estimation of Prediction Error: Covariance Penalties and Cross-Validation (with discussion),” *Journal of the American Statistical Association*, 99, 619–642.
- Golub, G., Heath, M., and Wahba, G. (1979), “Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter,” *Technometrics*, 21, 215–223.
- Hartshorne, R. (1977), *Algebraic Geometry*, Springer-Verlag, New York.
- Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall, London.

- Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.
- Hocking, R. and Leslie, R. (1967), “Selection of the Best Subset in Regression Analysis,” *Technometrics*, 9, 531–540.
- Hotelling, H. (1935), “The most predictable criterion,” *Journal of Educational Psychology*, 26, 139–142.
- Izenman, A. (1975), “Reduced-Rank Regression for the Multivariate Linear Model,” *Journal of Multivariate Statistics*, 5, 248–264.
- (2008), *Modern Multivariate. Statistical Techniques: Regression, Classification and Manifold Learning*, Springer, New York.
- Laub, A. (2004), *Matrix Analysis for Scientists and Engineers*, Society for Industrial and Applied Mathematics, Philadelphia.
- Lee, W. and Liu, Y. (2012), “Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood,” *Journal of Multivariate Analysis*, To Appear.
- Li, Y. and Zhu, J. (2007), “Analysis of array CGH data for cancer studies using the fused quantile regression,” *Bioinformatics*, 23, 2470–2476.
- Magnus, J. and Neudecker, H. (1998), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, New York.
- Mallow, C. (1973), “Some Comments on C_p ,” *Technometrics*, 15, 661–675.
- Massy, W. (1965), “Principal Component Regression with Exploratory Statistical Research,” *Journal of the American Statistical Association*, 60, 234–246.
- Meyer, M. and Woodroffe, M. (2000), “On the Degrees of Freedom in Shape-Restricted Regression,” *Annals of Statistics*, 28, 1083–1104.
- Mukherjee, A. and Zhu, J. (2011), “Reduced Rank Ridge Regression and its Kernel Extensions,” *Statistical Analysis and Data Mining*, 4, 612–622.
- Obozinski, G., Wainwright, M., and Jordan, M. (2011), “Support union recovery in high-dimensional multivariate regression,” *Annals of Statistics*, 39, 1–17.

- O’Neil, K. (2005), “Critical Points of the Singular Value Decomposition,” *SIAM Journal of Matrix Analysis and Applications*, 27, 459–473.
- Papadopoulos, T. and Lourakis, M. (2000), “Estimating the Jacobian of the Singular Value Decomposition: Theory and Applications,” in *In Proceedings of European Conference on Computer Vision, ECCV’00*, Springer, pp. 554–570.
- Peng, J., Zhu, J. Bergamaschi, A., Han, W., Noh, D., Pollack, J., and Wang, P. (2009), “Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer,” *Annals of Applied Statistics*, 4, 53–77.
- Rao, C. (1978), “Matrix Approximations and Reduction of Dimensionality in Multivariate Statistical Analysis,” in *In Proceedings of the Fifth International Symposium of Multivariate Analysis*, ed. Krishnaiah, P. R., North-Holland Publishing, pp. 3–22.
- Reinsel, G. and Velu, R. (1998), *Multivariate Reduced-Rank Regression: Theory and Applications*, Springer, New York.
- Rothman, A., Levina, L., and Zhu, J. (2009), “Sparse multivariate regression with covariance estimation,” *Journal of Computational and Graphical Statistics*, 19, 947–962.
- Shen, X. and Ye, J. (2002), “Adaptive Model Selection,” *Journal of the American Statistical Association*, 97, 210–221.
- Stein, C. (1981), “Estimation of the Mean of a Multivariate Normal Distribution,” *Annals of Statistics*, 9, 1135–1151.
- Tibshirani, R. (1996), “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society*, 58, 267–288.
- Turlach, B., Venables, W., and Wright, S. (2005), “Simultaneous variable selection,” *Technometrics*, 47, 349–363.
- Wille, A., Zimmermann, P., Vranova, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Bühlmann, P. (2004), “Sparse Graphical Gaussian Modeling of the Isoprenoid Gene Network in Arabidopsis Thaliana,” *Genome Biology*, 5, 1–13.
- Witten, D., Tibshirani, R., and Hastie, T. (2009), “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis,” *Biostatistics*, 10, 515–534.

- Wold, H. (1975), “Soft modeling by latent variables: the non-linear iterative partial least squares approach,” *Perspectives in Probability and Statistics, papers in honour of M. S. Bartlett*.
- Wright, K. (1992), “Differential Equations for the Analytic Singular Value Decomposition of a Matrix,” *Numerische Mathematik*, 63, 283–295.
- Ye, J. (1998), “On Measuring and Correcting the Effects of Data Mining and Model Selection,” *Journal of American Statistical Association*, 93, 120–131.
- Yuan, M., Ekici, L., Lu, Z., and Monteiro, R. (2007), “Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression,” *Journal of the royal Statistical Society*, 69, 329–346.
- Zou, H., Hastie, T., and Tibshirani, R. (2007), “On the “Degrees of Freedom” of the Lasso,” *Annals of Statistics*, 35, 2173–2192.

Appendix

Derivation of (8)

Using simple matrix calculus, trace identity $tr(\mathbf{AB}) = tr(\mathbf{BA})$ and (7), we get the following simplified expression of (8):

$$\begin{aligned}
\hat{df}(r) &= tr \left\{ \frac{\partial vec(\hat{\mathbf{Y}}(r))}{\partial vec(\mathbf{Y})} \right\} \\
&= tr \left\{ [\mathbf{I}_q \otimes \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1/2}] \left(\frac{\partial vec(\mathbf{W}^r \mathbf{D}^r \mathbf{V}^{rT})}{\partial vec(\mathbf{Y})} \right) \right\} \\
&= tr \left\{ [\mathbf{I}_q \otimes \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1/2}] \left(\frac{\partial vec(\mathbf{W}^r \mathbf{D}^r \mathbf{V}^{rT})}{\partial vec(\mathbf{F})} \right) \cdot \left(\frac{\partial vec(\mathbf{F})}{\partial vec(\mathbf{Y})} \right) \right\} \\
&= tr \left\{ [\mathbf{I}_q \otimes \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1/2}] \left(\frac{\partial vec(\mathbf{W}^r \mathbf{D}^r \mathbf{V}^{rT})}{\partial vec(\mathbf{F})} \right) [\mathbf{I}_q \otimes (\mathbf{X}^T \mathbf{X})^{-1/2} \mathbf{X}^T] \right\} \\
&= tr \left\{ \frac{\partial vec(\mathbf{W}^r \mathbf{D}^r \mathbf{V}^{rT})}{\partial vec(\mathbf{F})} \right\}_{pq \times pq}.
\end{aligned}$$